Hypothesis Testing and P-Values

A student-friendly introduction to statistical significance, decision rules, and interpretation in research

UNIVERSITY LECTURE

Author: Muhammad Hassan
Website: ResearchMethod.net

Lecture Overview

What this session covers and how the ideas connect

Concepts

Null and alternative hypotheses, alpha, p-values, test statistics

Process

How researchers move from a research question to an inferential decision

Interpretation

What p-values mean, what they do not mean, and how to report results

Application

Examples from social science, education, health, business, psychology, and IT

Quality

Best practices, ethics, common mistakes, and classroom exercises

Learning Objectives

By the end of this lecture, students should be able to:

1. Define hypothesis testing, null hypothesis, alternative hypothesis, alpha level, test statistic, and p-value.
2. Explain the logical steps used in a statistical hypothesis test.
3. Distinguish one-tailed and two-tailed tests and select an appropriate test direction.
4. Interpret p-values accurately without confusing them with effect size or practical importance.
5. Identify Type I error, Type II error, and statistical power in research decisions.
6. Choose common statistical tests for typical research questions and data types.
7. Report hypothesis testing results clearly, ethically, and transparently.

Introduction: Why Hypothesis Testing Matters

Hypothesis testing helps researchers evaluate whether sample evidence is strong enough to challenge a default explanation.

In Research

Transforms research questions into testable statistical statements.

Supports decisions when researchers only have sample data.

Helps separate random variation from patterns worth further attention.

Used across experiments, surveys, observational studies, and program evaluations.

In Practice

A lecturer tests whether a new teaching strategy improves exam scores.

A hospital checks whether a treatment changes recovery time.

A business compares customer satisfaction before and after a service change.

A psychologist examines whether stress scores differ between groups.

Key Definitions

Core vocabulary used throughout hypothesis testing

Hypothesis

A testable statement or prediction about a population parameter or relationship.

Null hypothesis (H_0)

The default claim, often stating no difference, no effect, or no relationship.

Alternative hypothesis (H_1/Ha)

The research claim that an effect, difference, or relationship exists.

P-value

The probability of data as extreme as observed, assuming H_0 and the test model are true.

Significance level (α)

The pre-set threshold for deciding whether evidence is statistically significant.

Test statistic

A standardized value that summarizes how far the sample result is from H_0.

Background and Context

Modern hypothesis testing grew from several statistical traditions

Fisherian significance testing

Used p-values as a measure of how incompatible the data are with a null model. The focus is evidence against H_0, not automatic accept/reject rules.

Neyman–Pearson framework

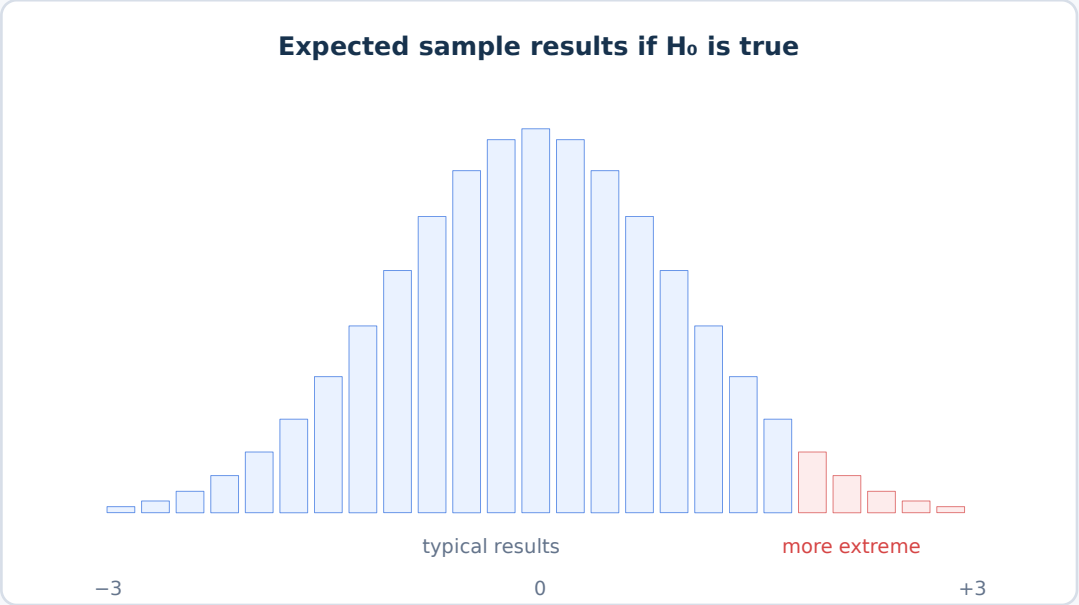
Introduced formal decision rules, long-run error rates, Type I and Type II errors, and statistical power.

Contemporary guidance

Researchers are encouraged to report p-values with effect sizes, confidence intervals, study design details, and transparent assumptions.

The Logic of Hypothesis Testing

A test asks: “Would this result be unusual if the null hypothesis were true?”



Start with a default explanation

Assume no real effect, no difference, or no relationship unless the data provide enough evidence.

Measure how unusual the sample result is

The test statistic places the observed result on a reference distribution.

Make a cautious inference

A small p-value suggests the observed result would be unlikely under H_0, but does not prove H_1.

Null and Alternative Hypotheses

Hypotheses convert research ideas into statistical statements

Null hypothesis (H_0)

Usually states “no effect,” “no difference,” or “no association.”

Represents the reference model used to calculate the p-value.

Is not “proved true” when a test is not significant.

Example: The mean exam score is the same for both teaching methods.

Alternative hypothesis (H_1 or Ha)

States the expected effect, difference, or association.

May be directional or non-directional.

Represents the research expectation being evaluated.

Example: The new teaching method changes the mean exam score.

Types of Hypotheses: One-Tailed and Two-Tailed Tests

The alternative hypothesis determines which results count as “extreme”

Two-tailed test

Used when the researcher expects a difference or relationship but does not specify direction.

Example: Students taught online and face-to-face have different mean scores.

Right-tailed test

Used when the researcher expects the value to be greater than the null value.

Example: A new program increases average attendance.

Left-tailed test

Used when the researcher expects the value to be less than the null value.

Example: A treatment reduces average recovery time.

Significance Level, Errors, and Power

Hypothesis testing balances evidence with the risk of wrong decisions

Decision / Reality	H_0 is true	H_0 is false
Reject H_0	Type I error False positive Probability = α	Correct decision Power = 1 − β
Do not reject H_0	Correct decision	Type II error False negative Probability = β

α (alpha)

The maximum acceptable risk of a Type I error, set before analysis.

β (beta)

The risk of missing a real effect or difference.

Power

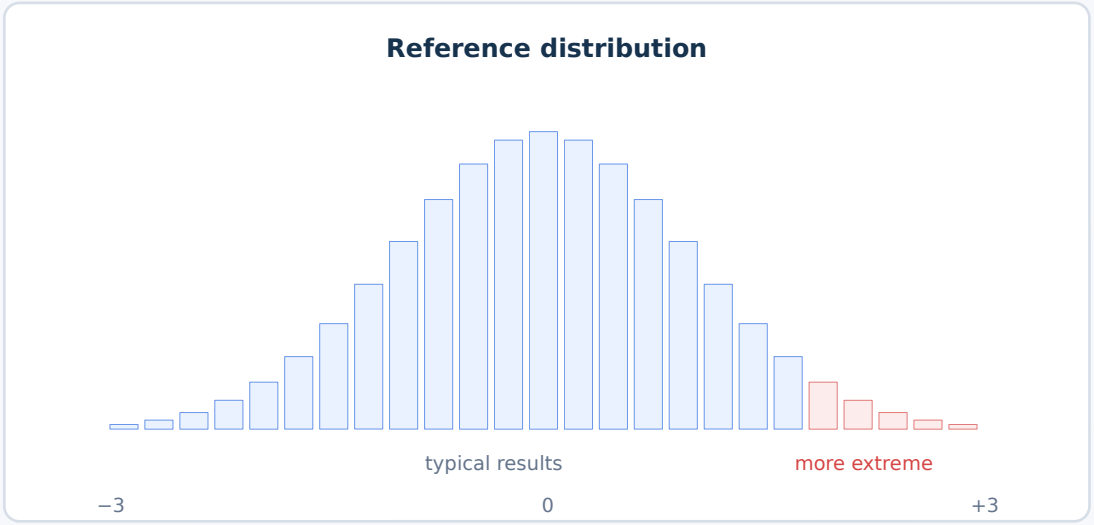
The probability of detecting a true effect when it exists.

Power = 1 − β

Test Statistics and Sampling Distributions

The test statistic converts a sample result into a standardized measure of distance from H_0

Test statistic = observed result − null value / standard error



Why standardization matters

Different studies use different units and sample sizes. A test statistic expresses the result in relation to expected sampling variation.

Common test statistics include z, t, χ^2, F, and regression coefficients divided by standard errors.

What Is a P-Value?

A p-value measures how surprising the observed data are under the null hypothesis

$$p = P(\text{data as extreme or more extreme} \mid H_0 \text{ and model assumptions})$$

Small p-value

The observed result would be relatively unusual if H_0 were true. This provides evidence against H_0.

Large p-value

The data are not unusual enough to reject H_0. This does not prove that H_0 is true.

Context matters

The p-value depends on sample size, variability, test choice, assumptions, and study design.

What a P-Value Does Not Mean

Many interpretation errors come from treating p-values as stronger than they are

Not the probability H_0 is true

p = .03 does not mean there is a 3% chance that the null hypothesis is true.

Not the probability results are due to chance alone

It is calculated under a model and cannot explain every source of uncertainty.

Not the size or importance of an effect

A tiny effect may be statistically significant in a very large sample.

Not a measure of study quality

Bias, poor measurement, and invalid assumptions can make any p-value misleading.

Not a replication guarantee

A significant result does not ensure that another study will obtain the same finding.

Step-by-Step Process for Hypothesis Testing

A disciplined sequence reduces bias and improves interpretation

Before looking at results

Define the outcome, groups or variables, hypotheses, alpha level, analysis plan, and exclusion rules. This protects against selective reporting.

After analysis

Report the p-value with the test used, sample size, effect size, confidence interval, assumptions, and practical meaning.

Common Statistical Tests: Choosing the Right Tool

The correct test depends on the research question, data type, and design

Research goal	Typical data	Common test	Example question
Compare one mean to a value	Continuous outcome	One-sample t-test	Is average satisfaction different from 3.5?
Compare two means	Continuous outcome, two groups	Independent-samples t-test	Do two teaching methods produce different scores?
Compare 3+ means	Continuous outcome, 3+ groups	ANOVA	Do scores differ across three programs?
Association between categories	Categorical variables	Chi-square test	Is gender associated with product preference?
Relationship between continuous variables	Continuous variables	Correlation / regression	Does study time predict exam score?

Tests for Means: One-Sample and Two-Sample Designs

Mean comparisons are among the most common applications of hypothesis testing

One-sample test

Compares a sample mean to a known or hypothesized population value.

Use when there is one group and a continuous outcome.

Example H_0: The average anxiety score equals 20.

Interpretation focuses on whether the sample mean is unusually far from the null value.

Two-sample test

Compares means between two independent groups or two paired measurements.

Independent: different participants in each group.

Paired: same participants measured twice or matched pairs.

Example H_0: Mean scores are equal for online and in-person students.

Tests for Categorical Data: Chi-Square Logic

Chi-square tests compare observed counts with expected counts under H_0

Preference	Observed	Expected under H_0
Product A	72	60
Product B	48	60
Total	120	120

$$\chi^2 = \Sigma \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

When to use

Use a chi-square goodness-of-fit test for one categorical variable, or a chi-square test of independence for two categorical variables.

The p-value tells whether the observed count pattern would be unusual under the null model.

Testing Relationships: Correlation and Regression

Hypothesis tests can evaluate whether variables are associated or predictive

Correlation

Tests whether two continuous variables move together.

H_0: $\rho = 0$

Example: Is study time related to exam score?

Regression coefficient

Tests whether a predictor explains variation in an outcome while accounting for model structure.

H_0: $\beta = 0$

Example: Does training hours predict productivity?

Important caution

Statistical association is not the same as causation. Causal claims need appropriate design, measurement, and control of confounding.

Practical Examples Across Disciplines

Hypothesis testing is used whenever researchers compare evidence with a null explanation

Social sciences

Do urban and rural voters differ in average trust in institutions?

Business

Did a loyalty program increase average monthly purchases?

Education

Do students using an online practice tool score higher on quizzes?

Health sciences

Does a medication reduce average blood pressure compared with standard care?

Psychology

Is mindfulness training associated with lower stress scores?

Information technology

Does a new interface reduce task-completion time?

Detailed Research Example

Scenario: Does formative feedback improve academic writing scores?

Research topic

Effect of structured formative feedback on undergraduate writing performance.

Research problem

Students often receive final grades but limited feedback before submission.

Research question

Do students receiving weekly feedback have higher final writing scores than students receiving usual instruction?

Participants and data

Two sections of first-year students. Outcome: final rubric score. Data collected after a 6-week writing module.

Analysis method

Independent-samples t-test comparing mean writing scores between the feedback and usual-instruction groups.

Expected interpretation

If $p < \alpha$, conclude that the data provide evidence of a difference in mean scores; also report effect size and confidence interval.

Interpreting and Reporting Results

Good reporting explains both the statistical result and the research meaning

Recommended reporting elements

Name the statistical test used.

Report sample size or degrees of freedom where relevant.

Report the test statistic and exact p-value when possible.

Include effect size and confidence interval.

Translate the result into the research context.

Mention key assumptions and any limitations.

Example statement

An independent-samples t-test indicated that the feedback group had higher mean writing scores than the usual-instruction group, t(df) = value, p = value. The difference should be interpreted alongside the effect size, confidence interval, and study design limitations.

Comparison Table: Related Concepts

P-values are only one part of statistical interpretation

Concept	Main question answered	What it does not answer	Best use
P-value	How unusual are the data under H_0?	Size, importance, or truth of H_0	Evidence against a null model
Alpha (α)	What false-positive risk is acceptable before testing?	How surprising the actual data are	Decision threshold
Effect size	How large is the effect or relationship?	Whether sampling error alone explains the result	Practical importance
Confidence interval	What range of values is compatible with the data?	A single yes/no decision by itself	Precision and uncertainty

Advantages of Hypothesis Testing

Why researchers continue to use hypothesis tests

Provides a structured and transparent approach to evaluating sample evidence.

Allows researchers to quantify uncertainty using a common statistical language.

Supports comparison across groups, time points, treatments, and predictors.

Encourages pre-specification of hypotheses, alpha levels, and analysis plans.

Works with many research designs and data types when assumptions are appropriate.

Helps communicate findings in academic articles, theses, reports, and presentations.

Limitations and Common Concerns

A statistically significant result is not automatically a strong or useful result

Dependence on assumptions

P-values can be misleading when sampling, independence, measurement, distributional, or model assumptions are violated.

Sensitivity to sample size

Very large samples may detect trivial effects; small samples may fail to detect meaningful effects.

Misuse and overreliance

Focusing only on $p < .05$ can hide uncertainty, effect magnitude, bias, data quality, and practical relevance.

Common Mistakes by Students and Beginning Researchers

Most problems are interpretation problems, not calculation problems

Saying “p = .04 means the null hypothesis has a 4% chance of being true.”

Treating p < .05 as proof of a theory or causal relationship.

Choosing one-tailed tests after seeing the direction of results.

Reporting only “significant” or “not significant” without effect sizes.

Ignoring assumptions, outliers, missing data, and measurement quality.

Running many tests and selectively reporting the smallest p-values.

Confusing statistical significance with practical, clinical, or educational importance.

Best Practices for Using P-Values

Good inference combines p-values with context, design quality, and transparency

Plan first

State hypotheses, analysis method, alpha level, exclusion rules, and variables before analyzing the data.

Report fully

Include test statistic, p-value, effect size, confidence interval, sample size, assumptions, and limitations.

Interpret cautiously

Discuss whether the result is meaningful in the research context and whether alternative explanations remain possible.

Ethical Considerations in Hypothesis Testing

Statistical decisions affect credibility, policy, and people

Pre-register hypotheses and analysis plans when feasible to reduce selective reporting.

Avoid p-hacking, HARKing, cherry-picking, and hiding non-significant findings.

Report methods, exclusions, missing data handling, and all relevant analyses transparently.

Protect participants’ privacy and confidentiality when reporting datasets and results.

Do not overstate causality when the study design is observational or weakly controlled.

Interpret results in ways that avoid stigmatizing groups or misusing statistical evidence.

Classroom Activity: Interpret the P-Value

A short group exercise for applying the lecture concepts

Activity objective

Practice writing hypotheses, selecting a decision rule, and interpreting a p-value without overclaiming.

Instructions

Groups receive this result: $p = .032$ for a two-tailed test comparing two teaching methods at $\alpha = .05$. Students write H_0, H_1, decision, and interpretation.

Expected outcome

Students should conclude that the result is statistically significant at $\alpha = .05$, but should avoid claiming proof, causation, or practical importance without more information.

Discussion Questions

Use these prompts to encourage critical thinking

1. Why is the null hypothesis used as the reference point for a p-value?
2. What is the difference between statistical significance and practical importance?
3. When might a non-significant result still be scientifically useful?
4. Why should researchers choose α before analyzing data?
5. How can large sample sizes affect p-values and interpretation?
6. What information should be reported alongside a p-value?
7. How can p-hacking damage the credibility of research?

Knowledge Check

Answer the questions individually, then discuss in pairs

1. What does H_0 usually represent in hypothesis testing?
2. True or false: A p-value is the probability that H_0 is true.
3. A researcher sets $\alpha = .05$ and obtains $p = .018$. What is the decision?
4. What is a Type I error?
5. Which test is commonly used to compare two independent group means?
6. Why should effect sizes be reported with p-values?
7. What is one ethical problem related to p-values?
8. What does a two-tailed test evaluate?

Knowledge Check: Answer Key

Suggested answers for instructors and students

1. The default claim of no effect, no difference, or no relationship.
2. False. It is conditional on H_0 and the statistical model being true.
3. Reject H_0 at $\alpha = .05$ because $p < \alpha$.
4. Rejecting H_0 when H_0 is actually true; a false positive.
5. Independent-samples t-test, assuming test conditions are appropriate.
6. Effect sizes show magnitude or practical importance, which p-values do not show.
7. Examples include p-hacking, selective reporting, hiding non-significant results, or overstating conclusions.
8. Whether the parameter differs from the null value in either direction.

Key Takeaways

The most important ideas to remember

1. Hypothesis tests evaluate evidence under a null model

They do not prove that a research hypothesis is true.

2. A p-value is conditional

It assumes the null hypothesis and model assumptions are correct.

3. Statistical significance is not enough

Effect size, confidence intervals, design quality, and practical context are essential.

4. Transparent reporting improves credibility

Pre-planning, assumption checks, and honest reporting reduce misuse.

5. Good researchers interpret results cautiously

A result should be discussed as evidence, not as absolute proof.

Conclusion

Hypothesis testing and p-values are powerful tools when used carefully

Hypothesis testing provides a formal way to compare observed sample evidence with what would be expected under a null hypothesis. P-values help summarize how unusual the data are under that model, but they must be interpreted with caution. Strong research conclusions require more than $p < .05$: they require sound design, valid measurement, transparent analysis, effect sizes, confidence intervals, and ethical reporting.

Final message

Use p-values as one piece of evidence in a broader research argument—not as a substitute for careful thinking.

References

Selected credible sources in APA 7th edition format

Agresti, A., & Finlay, B. (2018). Statistical methods for the social sciences (5th ed.). Pearson.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Field, A. (2018). Discovering statistics using IBM SPSS Statistics (5th ed.). SAGE Publications.

Fisher, R. A. (1925). Statistical methods for research workers. Oliver and Boyd.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London. Series A, 231, 289–337. https://doi.org/10.1098/rsta.1933.0009

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. The American Statistician, 70(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond “p < 0.05”. The American Statistician, 73(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Thank You

Author: Muhammad Hassan
Website: ResearchMethod.net

Visit ResearchMethod.net for research guides, templates, lecture slides, examples, and academic resources.